

**Manuscript version: Author's Accepted Manuscript**

The version presented in WRAP is the author's accepted manuscript and may differ from the published version or Version of Record.

**Persistent WRAP URL:**

<http://wrap.warwick.ac.uk/132624>

**How to cite:**

Please refer to published version for the most recent bibliographic citation information. If a published version is known of, the repository item page linked to above, will contain details on accessing it.

**Copyright and reuse:**

The Warwick Research Archive Portal (WRAP) makes this work by researchers of the University of Warwick available open access under the following conditions.

Copyright © and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable the material made available in WRAP has been checked for eligibility before being made available.

Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

**Publisher's statement:**

Please refer to the repository item page, publisher's statement section, for further information.

For more information, please contact the WRAP Team at: [wrap@warwick.ac.uk](mailto:wrap@warwick.ac.uk).

# Heap: A command for estimating discrete outcome variable models in the presence of heaping at known points

Zizhong Yan  
Jinan University

Wiji Arulampalam  
University of Warwick  
Daniel Gutknecht  
Goethe University Frankfurt

Valentina Corradi  
University of Surrey

**Abstract.** Self-reported survey data are often plagued by the presence of heaping. Accounting for this measurement error is crucial for the identification and consistent estimation of the underlying model (parameters) from such data. This paper introduces two **Stata** commands. The first command, **heapmph**, estimates the parameters of a discrete-time mixed proportional hazard model with gamma unobserved heterogeneity, allowing for fixed and individual-specific censoring, and different sized heap points. The second command, **heapop**, extends the framework to ordered choice outcomes, subject to heaping. Suitable specification tests are also provided.

**Keywords:** st0001, **heapmph**, **heapop**, Discrete time duration model, Mixed proportional hazards model, Ordered choice model, Heaping, Measurement Error.

**Acknowledgements:** We are grateful to the British Academy (grant number: SG160731 - Estimation and inference with heaped data - a novel approach), for funding this project. Zizhong Yan acknowledges the support from the 111 project of China (project number B18026). We also thank David M. Drukker and a referee of this journal for helpful comments and discussions.

## 1 Introduction

A problem frequently encountered in survey data is the abnormal concentration of reported observations at certain values of the outcome variable. Examples include reported dates of death in neo-natal mortality data (Arulampalam et al. 2017, ACG from now on), age of starting and quitting cigarette smoking (Forster and Jones 2001), or self-reported consumption expenditure data (Pudney 2008). One of the main reasons for such concentration, often referred to as heap points, is rounding. Correctly identifying and accounting for the rounding behavior is crucial for consistent estimation of and valid inference on the parameters of the underlying model of interest. The paper ACG discusses identification and estimation of popular duration and ordered choice models, in the presence of heaping, using maximum likelihood procedures.

In this paper, we introduce the **Stata** command **heapmph** to estimate the underlying parameters in the case of a discrete-time mixed proportional hazard (Cox 1972) duration model as proposed in ACG. More specifically, this command estimates a semi-

parametric baseline hazard function in the presence of heaping of observations at certain durations, and gamma distributed unobserved heterogeneity (frailty). In the accompanying **heapop**, we extend the framework to an ordered choice model, allowing for the presence of heaping points.

As shown in ACG, when some of the parameters lie on the boundary of the parameter space, the limiting distribution of the estimator is no longer a normal distribution, and more complicated subsampling procedures are required for inference. Hence, we also provide two specification tests. The first one tests for the absence of heaping effects in the model. The second specification test examines whether all heaping parameters lie inside the parameter space, which in turn will allow for inference based on asymptotic normality. We use the so called  $M$  out of  $N$  bootstrap method to calculate the standard errors. These tests provide a set of tools that enable applied researchers to verify the validity of different model specifications.

In Appendix 1, we show how the **heapmph** command can be used to test for a shift in the heaping probability and/ or baseline parameters as a consequence of a policy or regime change, while in Appendix 2 we outline similar examples for **heapop**. Finally, in Appendix 3 we formally link the proportional hazard model to the Type I Extreme-Value (EV) ordered choice model (Han and Hausman 1990) outlining the implications for the interpretation of the parameters.

## 2 Mixed proportional hazard model with ‘heaping’

### 2.1 Specification

We start with the Mixed Proportional Hazard (MPH) model for the unobserved true durations in continuous time, and parameterize this for individual  $i$  as:

$$\lambda_i(\tau^*|z_i, u_i) = \lambda_0(\tau^*) \exp(z_i' \beta + u_i), \quad (1)$$

where  $\lambda_0(\tau^*)$  is the baseline hazard at time  $\tau^*$ ,  $u_i$  is the individual unobserved heterogeneity (frailty), and  $z_i$  a set of time invariant covariates. In most empirical studies, time is observed on a discrete scale. We therefore, assume that a continuous duration  $\tau_i^* \in [\tau, \tau + 1)$  is recorded as  $\tau$ , where  $\tau$  denotes a discrete time period, so that the sample of (discrete) durations is given by  $\tau_i$  for  $i = 1, \dots, N$ . The discrete time hazard for our model can then be written as:

$$\begin{aligned} h_i(\tau|z_i, u_i) &= \Pr[\tau_i^* < \tau + 1 | \tau_i^* \geq \tau, z_i, u_i] \\ &= 1 - \exp\left(-\int_{\tau}^{\tau+1} \lambda_i(s|z_i, u_i) ds\right) \\ &= 1 - \exp\left(-\exp\left(z_i' \beta + \gamma(\tau) + u_i\right)\right), \end{aligned} \quad (2)$$

where  $\gamma(\tau) = \ln \int_{\tau}^{\tau+1} \lambda_0(s) ds$ . Due to misreporting, the researcher however, does not observe  $\tau_i$  directly, but  $t_i$ , a potentially mismeasured version of it.

More specifically, the form of misreporting we address is referred to as “heaping” in the literature, and describes the phenomenon of observing an over- and under-reporting of failures at certain time periods. We briefly list informally the set of assumptions for the derivation of the estimator and its properties here, and refer the readers to ACG for further details on the assumptions and identification results.<sup>1</sup> Based on the neonatal mortality illustration from ACG, we also illustrate our command using a simulated dataset based on ACG.

### Assumptions:

- A1 Excessive concentrations of reported failures occur at time periods that are multiples of a positive integer. This implies equal distance between the heap points. In most of the empirical applications where we see heaping due to rounding, we often see the distance between heaping points to be the same. This is the scenario `heapmph` uses.<sup>2</sup> There is no heaping at time zero. This is not an unrealistic assumption, since one would expect survey respondents to know whether the discretized duration was a zero or not. Following ACG, our illustration also assumes the heaping to be at points that are multiples of 5.
- A2 In order to identify the baseline hazard from possibly misreported observations, we need to impose a structure on the heaping process. In the illustration provided here, we assume that one period to the right and to the left of each heap point are associated with that heap. We denote the maximum number of time periods that a duration can be rounded to as  $\bar{r}$ , and in this example  $\bar{r} = 1$ . That is, we assume that the duration points 4, 9, and 14, will be rounded up, while 6, 11, and 16 will be rounded down to 5, 10 and 15, respectively.
- A3 All heaping is to observed duration points only. In our example, this implies that the heaping is to the points 5, 10, and 15 only, as we assume that the outcome variable is censored at 18 days. The maximum number of heaps is assumed to be  $\bar{j}$ , and in our example  $\bar{j} = 3$ .
- A4 The censoring is exogenous, and the censored observations are correctly reported.
- A5 Whenever the true duration falls onto one of the heaping points, it will be correctly reported. However, whenever the duration falls onto the non-heaping points, it is assumed to be either correctly reported or rounded (up or down) to the nearest heaping point. Let  $p_1, p_2$ , etc. denote the corresponding rounding up probabilities when a true duration is lower by one, two, etc units from the nearest heaping point. Similarly let  $q_1, q_2$ , etc. denote the rounding down probabilities when a true duration is higher by one, two, etc. units from the nearest heaping point. In our illustration, a reported duration of say 10 days, includes true durations of 11 (9) days, which have been rounded down (up) to 10 days (see Figure 1). Hence,

1. Note, ACG discusses a more general setup which can accommodate more complex heaping mechanisms.

2. It is noteworthy that the theoretical setup can in principle be straightforwardly amended to allow for non-equally spaced heaping points, see the paper ACG.

$p_1$  is the probability that a true duration of 9 will be rounded up to 10 days. Analogously  $q_1$  is the probability that a true duration of 11 will be rounded down to 10 days.

- A6 There exists a segment in the baseline hazard that is constant from time period  $\bar{k}$ , and includes a known true value (i.e. there is no mis-reporting at this value). In our example, we assume  $\bar{k} = 12$ .

Heuristically, the assumption that the hazard is constant over a set of time periods, which includes (at least) a known true value, enables us to uniquely identify the  $\gamma$  parameter associated with this correctly reported time period as well as the parameters of the heaping process, i.e. the  $p$ s and the  $q$ s, in this region, from the observed data. Subsequently, we can use these identified probability parameters to pin down the rest of the baseline and other hazard parameters. See Figure 1.

## 2.2 Maximum likelihood estimation

Before writing down our likelihood function, we first define some notation.

Let  $\underline{\theta} = \{\beta', \gamma'\}'$  with  $\gamma = \{\gamma(0), \gamma(1), \dots, \gamma(\bar{\tau} - 1)\}'$ , and  $\bar{\tau}$  be some finite, positive integer, and  $(\bar{\tau} - 1)$  represent the uncensored maximum observed time period. Define the probability of survival at least until time period  $\tau < \bar{\tau}$  in the absence of misreporting as:

$$\begin{aligned} S_i(\tau|z_i, u_i, \underline{\theta}) &= \Pr(\tau_i \geq \tau|z_i, u_i, \underline{\theta}) \\ &= \prod_{s=0}^{\tau-1} \exp(-\exp(z_i' \beta + \gamma(s) + u_i)) \\ &= \prod_{s=0}^{\tau-1} \exp(-v_i \exp(z_i' \beta + \gamma(s))), \end{aligned}$$

where  $v_i \equiv \exp(u_i)$ , and  $u_i$  is the unobserved heterogeneity.

The probability for an exit event in  $\tau_i < \bar{\tau}$  is:

$$\begin{aligned} f_i(\tau|z_i, u_i, \underline{\theta}) &= \Pr(\tau_i = \tau|z_i, u_i, \underline{\theta}) \\ &= S_i(\tau|z_i, u_i, \underline{\theta}) - S_i(\tau + 1|z_i, u_i, \underline{\theta}) \\ &= \prod_{s=0}^{\tau-1} \exp(-v_i \exp(z_i' \beta + \gamma(s))) \\ &\quad - \prod_{s=0}^{\tau} \exp(-v_i \exp(z_i' \beta + \gamma(s))). \end{aligned} \tag{3}$$

$f_i(\tau|z_i, u_i, \underline{\theta})$  in the above equation denotes the probability of a duration equal to  $\tau$  when there is no misreporting. However, because of the rounding, heaped values are

over-reported while non-heaped values are under-reported, and this needs to be taken into account when constructing the likelihood function (see below).

Henceforth, let

$$\phi_i(t|z_i, v_i, \underline{\theta}) = \Pr(t_i = t|z_i, v_i, \underline{\theta})$$

with  $t_i$  denoting the discrete *reported* duration.

The likelihood contributions depend on the following four cases.

(I) For correctly reported durations,  $\phi_i(t|z_i, v_i, \underline{\theta}) = f_i(t|z_i, v_i, \underline{\theta})$ . This will include the duration point discussed in Assumption A3 earlier. Depending on the application, there might be other points too.

(II) For reported durations that are  $l = 1, 2$ , etc points *below* the nearest heaping point,  $\phi_i(t|z_i, v_i, \underline{\theta}) = (1 - p_l)f_i(t|z_i, v_i, \underline{\theta})$ , since  $p_l$  refer to the probabilities of rounding up.

(III) Similar to (II), for reported durations that are  $l = 1, 2$ , etc points *above* the nearest heaping point,  $\phi_i(t|z_i, v_i, \underline{\theta}) = (1 - q_l)f_i(t|z_i, v_i, \underline{\theta})$ , since  $q_l$  refer to the probabilities of rounding down.

(IV) Finally for reported durations on the heaping points:

$$\phi_i(t|z_i, v_i, \underline{\theta}) = \sum_l p_l f_i(t - l|z_i, v_i, \underline{\theta}) + \sum_l q_l f_i(t + l|z_i, v_i, \underline{\theta}) + f_i(t|z_i, v_i, \underline{\theta}).$$

In summary, there are four different probabilities of exit events depending on the nature of the true duration.

We next write down the corresponding unconditional probabilities under a set of assumptions on the unobserved heterogeneity  $v_i$ . More specifically, we impose the following assumptions on the properties and the distributional form of  $v_i$ , which are standard in the duration literature:

- (i)  $v_i$  is identically and independently distributed over  $i$  and is also independent of  $z_i$ ;
- (ii)  $v_i$  follows a Gamma distribution with unit mean and variance  $\sigma^2$ .<sup>3</sup>

The unconditional probabilities under the above assumptions, in case (I) above are

---

3. The assumption of Gamma distribution for  $v_i$  gives us a closed form expression for the unconditional probabilities. While the choice of the Gamma distribution might appear overly restrictive at first sight, we note that this can often be rationalized theoretically (Abbring and Van Den Berg 2007). In addition, findings by Han and Hausman (1990) as well as Meyer (1990) suggest that estimation results for discrete-time proportional hazard models where the baseline is left unspecified, display little sensitivity to alternative distributional assumptions.

given by:

$$\begin{aligned}
\int \phi_i(t|z_i, v, \underline{\theta}) g(v; \sigma) dv &= \int \Pr(\tau_i = t|z_i, v, \underline{\theta}) g(v; \sigma) dv \\
&= \int S_i(t|z_i, v, \underline{\theta}) g(v; \sigma) dv - \int S_i(t+1|z_i, v, \underline{\theta}) g(v; \sigma) dv \\
&= \left( 1 + \sigma \left( \sum_{s=0}^{t-1} \exp(z'_i \beta + \gamma(s)) \right) \right)^{-\sigma^{-1}} \\
&\quad - \left( 1 + \sigma \left( \sum_{s=0}^t \exp(z'_i \beta + \gamma(s)) \right) \right)^{-\sigma^{-1}}
\end{aligned}$$

where the last equality uses the fact that there is a closed form expression under the Gamma density assumption for  $v$  (e.g., see Meyer (1990, p. 770)). Moreover, since the integral is a linear operator, the probabilities for the cases (II) to (IV) can be derived accordingly.

Our next goal is to obtain consistent estimators for  $\theta = \{\underline{\theta}', \sigma, p_1, \dots, p_{\bar{r}}, q_1, \dots, q_{\bar{r}}\}'$  from the possibly misreported durations. Before setting up the likelihood function, we introduce censoring into our setup.

Let  $c_i$  be an indicator equal to one if the observation is uncensored and zero otherwise. It is assumed that durations are censored at a fixed time  $\bar{r}$  which exceeds the points that are rounded and is not one of the heaping points. Assuming that censoring is independent of the heaping process and the durations, we have the following unconditional likelihood contributions.<sup>4</sup>

The likelihood function for the observed sample is:

$$L_N(\theta) = \prod_{i=1}^N \int \left\{ \phi_i(t|z_i, v)^{c_i} S_i(t|z_i, v)^{(1-c_i)} \right\} g(v; \sigma) dv$$

and so

$$l_N(\theta) = \ln L_N(\theta) = \sum_{i=1}^N \ln \int \left\{ \phi_i(t|z_i, v)^{c_i} S_i(t|z_i, v)^{(1-c_i)} \right\} g(v; \sigma) dv.$$

Given the definition of  $\phi_i(t|z_i, v)$  and cases (I) through (IV), it is clear that the (log) likelihood function down-weights the contribution of heaped durations, and over-weights the contribution of non heaped durations.

Under the assumptions provided in ACG, it can be shown that the limiting distribution of the estimator depends on whether some heaping probability parameters lie on the boundary of the parameter space or not, that is, if one or more of the “true”

---

4. For ease of exposition, we have assumed a constant censoring point (type I censoring; Cox and Oakes 1984). However, the program allows the censoring points to vary over  $i$ .

probability parameters are equal to zero. In this case, the limiting distribution is no longer normal as the information matrix is not block diagonal in general, but takes a different form. We use the  $M$  out of  $N$  bootstrap method to derive the asymptotic standard errors. Details are provided in ACG.

### 3 Ordered probit model with heaping: specification and estimation

In general, there are many observed discrete outcomes (other than durations) that can exhibit heaping. For instance, survey data on the number of doctor visits or on cigarette consumption in a given period of time is often subject to this phenomenon. Here we discuss the estimation of an ordered probit model allowing for heaping. In Appendix 3, we provide a discussion on the link between the discrete duration model derived from the proportional hazard specification and the ordered choice model. To keep notational clutter to a minimum, we do not explicitly show the conditioning set in what follows.

Consider the following latent variable model representation of an ordered choice model.<sup>5</sup>

$$y_i^* = z_i' \beta^\dagger + \varepsilon_i$$

where  $y_i^*$  represents the latent outcome,  $z_i$  stands for the vector of regressors,  $\beta^\dagger$  is the vector of coefficients, and let the cumulative probability function of the error term  $\varepsilon_i$  be standard normal, denoted by  $\Phi(\cdot)$ .<sup>6</sup> Assume we have an ordered discrete outcome variable coded as  $y_i \in \{0, \dots, J\}$ . That is, we have:

$$y_i = j \text{ if and only if } \kappa_j < y_i^* = z_i' \beta^\dagger + \varepsilon_i < \kappa_{j+1},$$

where  $\kappa_0, \dots, \kappa_J$  are the threshold parameters that divide the real line into a finite number of intervals. Here, we have assumed the normalizations  $\kappa_0 = -\infty$ ,  $\kappa_{J+1} = +\infty$ , and  $\kappa_j < \kappa_{j+1}$ . In addition, note that we require a scale normalization and so  $z_i$  may not contain a constant. For any  $j \in \{0, \dots, J\}$ , the probabilities of interest are given by:

$$\begin{aligned} \Pr(y_i = j) &= \Pr(\kappa_j < y_i^* < \kappa_{j+1}) \\ &= \Pr(\kappa_j - z_i' \beta^\dagger < \varepsilon_i < \kappa_{j+1} - z_i' \beta^\dagger) \\ &= \Phi(\kappa_{j+1} - z_i' \beta^\dagger) - \Phi(\kappa_j - z_i' \beta^\dagger). \end{aligned} \quad (4)$$

In the presence the heaping data, the term  $\Pr(y_i = j)$  depends on the four cases:

(I) For correctly reported outcomes,  $\Pr(y_i = j) = \Phi(\kappa_{j+1} - z_i' \beta^\dagger) - \Phi(\kappa_j - z_i' \beta^\dagger)$ .

5. Supplementary material provided in ACG sketches the key identification conditions required for the estimation of this model when heaping is present in the data. A class of ordered choice models known as generalized ordered choice models, extends the standard model in different ways to incorporate unobserved heterogeneity (Greene 2014). Our `Stata` command estimates the standard ordered probit model with heaping, but without unobserved heterogeneity.

6. In principle, other distributions for  $\varepsilon_i$  can be chosen. For example, a logistic distribution for  $\varepsilon_i$  will lead to an ordered logit model.



(II) For reported outcomes that are  $l = 1, 2$ , etc. points *below* the nearest heaping point,  $\Pr(y_i = j) = (1 - p_l)(\Phi(\kappa_{j+1} - z'_i\beta^\dagger) - \Phi(\kappa_j - z'_i\beta^\dagger))$ .

(III) Similar to (II), for reported outcomes that are  $l = 1, 2$ , etc. points *above* the nearest heaping point,  $\Pr(y_i = j) = (1 - q_l)(\Phi(\kappa_{j+1} - z'_i\beta^\dagger) - \Phi(\kappa_j - z'_i\beta^\dagger))$ .

(IV) Finally for reported outcomes on the heaping points:

$$\begin{aligned} \Pr(y_i = j) &= \sum_l p_l (\Phi(\kappa_{j+1} - z'_i\beta^\dagger) - \Phi(\kappa_j - z'_i\beta^\dagger)) \\ &\quad + \sum_l q_l (\Phi(\kappa_{j+1} - z'_i\beta^\dagger) - \Phi(\kappa_j - z'_i\beta^\dagger)) \\ &\quad + (\Phi(\kappa_{j+1} - z'_i\beta^\dagger) - \Phi(\kappa_j - z'_i\beta^\dagger)) \end{aligned}$$

Note that when the outcome is duration data and for right-censored data at  $y_i = \bar{\tau}$ , the likelihood function can be written as:

$$L_N(\theta^\dagger) = \sum_{i=1}^N \left( \sum_{j=1}^{\bar{\tau}-1} \Pr(y_i = j) \right)^{d_{ij} \cdot c_i} (1 - \Phi(\kappa_{\bar{\tau}} - z'_i\beta^\dagger))^{(1-c_i)}, \quad (5)$$

where  $\theta^\dagger = \{\beta^\dagger, \kappa', p_1, \dots, p_{\bar{\tau}}, q_1, \dots, q_{\bar{\tau}}\}'$  and  $d_{ij}$  is an indicator equal to one when  $t_i = j$  and zero otherwise.

## 4 Testing for ‘heaping’

As pointed out in Section 2.2, if some of the heaping probability parameters lie on the boundary of the parameter space, the asymptotic distribution of the estimator is no longer normal. In addition, inference becomes more complicated, since subsampling methods are used to derive the asymptotic standard errors. In the following, we discuss two specification tests. First, a test to detect whether heaping matters in a statistical sense ( $\mathbf{H}^{\pi_1}$ ). If heaping matters, a second test to discriminate between the general case that allows for probability parameters on the boundary, and the special case without parameters on the boundary ( $\mathbf{H}^{\pi_2}$ ). That is, while the first test helps to determine whether the specified heaping model is indeed preferred over a standard model that does not account for heaping, the second test allows one to decide whether inference, in fact, ought to be based on subsampling methods.

Thus, collecting all heaping parameters in the vector  $\pi$  with  $\pi = \{p_1, \dots, p_{\bar{\tau}}, q_1, \dots, q_{\bar{\tau}}\}'$  and  $\theta = \{\theta', \sigma, \pi'\}'$ , the first test examines the existence of heaping effects through:

$\mathbf{H}^{\pi_1}$ :

$$H_0^{\pi_1} : p_1 = \dots = p_{\bar{\tau}} = q_1 = \dots = q_{\bar{\tau}} = 0$$

vs

$$H_A^{\pi_1} : p_l > 0 \text{ and/or } q_l > 0$$

for some  $l = 1, \dots, \bar{r}$ . The above hypothesis  $H_0^{\pi_1}$  can be tested through a standard likelihood ratio test (ACG).

The second specification test examines whether all heaping parameters lie inside the parameter space, which in turn allows inference based on asymptotic normality. That is, the null hypothesis of the test is that at least one rounding parameter is equal to zero versus the alternative that none is zero (and thus no boundary problem exists). Therefore, if we reject this hypothesis, we are able to make inference based on standard normal critical values, while if we fail to reject we ought to rely on subsampling methods for inference.

Formally, let  $H_{p,0}^{(j)} : p_j = 0$ ,  $H_{p,A}^{(j)} : p_j > 0$ , and let  $H_{q,0}^{(j)}, H_{q,A}^{(j)}$  be defined analogously. Our objective is to test the following hypothesis:

$H^{\pi_2}$ :

$$H_0^{\pi_2} = \left( \bigcup_{j=1}^{\bar{r}} H_{p,0}^{(j)} \right) \cup \left( \bigcup_{j=1}^{\bar{r}} H_{q,0}^{(j)} \right)$$

vs

$$H_A^{\pi_2} = \left( \bigcap_{j=1}^{\bar{r}} H_{p,A}^{(j)} \right) \cap \left( \bigcap_{j=1}^{\bar{r}} H_{q,A}^{(j)} \right),$$

so that under  $H_A^{\pi_2}$  all  $p$ s and  $q$ s are strictly positive. To discriminate between  $H_0^{\pi_2}$  and  $H_A^{\pi_2}$ , we apply the Intersection-Union principle (IUP), see e.g. chapter 5 in Silvapulle and Sen (2005). According to the IUP, we only reject  $H_0^{\pi_2}$  at level  $\alpha$  if all single null hypotheses  $H_{p,0}^{(j)}$  and  $H_{q,0}^{(j)}$  are rejected at level  $\alpha$ .

We now introduce a rule to discriminate between  $H_0^{\pi_2}$  and  $H_A^{\pi_2}$ .

**Rule IUP-PQ:** Reject  $H_0^{\pi_2}$ , if  $\max_{j=1, \dots, \bar{r}} \{PV_{p,j}, PV_{q,j}\} < \alpha$  and, do not reject otherwise.

Thus, as pointed out above, if one rejects  $H_0^{\pi_2}$ , the inference can be based on asymptotic normality, while failure to reject  $H_0^{\pi_2}$  requires the use of subsampling methods as outlined before.

## 5 Command Implementation

As discussed in the earlier section, if one or more of the probability parameters lie on the boundary of the parameter space, the asymptotic distribution of the estimator is no longer normal. We provide two tests that can be used to detect this. Hence, the output provides the usual asymptotic standard errors along with the standard errors calculated using the  $M$  out of  $N$  bootstrap method, where  $M$  denotes an integer strictly smaller than  $N$  (see ACG).

### 5.1 Data

We illustrate the use of the `heapmph` and `heapop` commands using generated data based on 250 observations drawn randomly from the original sample used in ACG. More

specifically, we retain two covariates of these observations that were found to be significant: mother's age at the time of birth (`age.m`), and mother's years of schooling (`school.m`). Our outcome variable `duration`, which is the time of death of the child measured in days if the child died within the first 17 days, is generated using these two covariates within the ordered choice model framework as detailed next. All observations where the child survived for longer than 18 days are treated as censored.<sup>7</sup>

The latent dependent variable  $y_i^*$  in the ordered choice model framework, is generated according to:

$$y_i^* = 0.1 \text{ age.m}_i - 0.1 \text{ school.m}_i + \varepsilon_i \quad \text{for } i = 1, \dots, 250.$$

We use two different schemes to generate  $\varepsilon_i$  for demonstrating `heapmph` and `heapop` commands, respectively. Note, as shown in Appendix 3, the Cox's proportional hazards (PH) model is equivalent to the ordered choice model where the underlying error term in the latent variable model is Type I EV distributed. The threshold parameters  $\kappa$  are then generated in terms of parameters  $\gamma$  (see Appendix 3).<sup>8</sup> In detail:

(i) For `heapmph` command, we characterize a proportional hazard model data example by generating i.i.d.  $\varepsilon_i$  from a Type I EV distribution. The baseline gamma parameters are set as follows:  $\exp(\gamma(t)) = 0.3$  for  $t = 0, 1, 2, 3$ ,  $\exp(\gamma(t)) = 0.6$  for  $t = 4, \dots, 7$ ,  $\exp(\gamma(t)) = 1.2$  for  $t = 8, \dots, 11$ ,  $\exp(\gamma(t)) = 2.5$  for  $t = 12, \dots, 15$ ,  $\exp(\gamma(16)) = 8$ , and  $\exp(\gamma(17)) = 10$ . The dataset created according to this scheme is enclosed in the package and named as "`heap.demonstration2.dta`".

(ii) For the data example used to demonstrating `heapop`, we draw  $\varepsilon_i$  from a standard normal distribution. We set the gamma parameters for `heapop` as follows:  $\exp(\gamma(t)) = 0.6$  for  $t = 0, 1, \dots, 11$ ,  $\exp(\gamma(t)) = 1.5$  for  $t = 12, \dots, 15$ ,  $\exp(\gamma(16)) = 1.8$ , and  $\exp(\gamma(17)) = 3$ . In the `heap` package, the dataset generated following this scheme is named as "`heap.demonstration.dta`".

Note that we keep the function flat from period 12 to 15. The discrete duration variable without heaping, for each observation  $i = 1, 2, \dots, 250$ , for these models is then generated using the cutoff points as:

$$\text{duration}_i = t \text{ if } y_i^* \in [\delta_t, \delta_{t+1}) \quad \text{for } t = 0, \dots, 18$$

where we assume  $\delta_0 = -\infty$ , and  $\delta_{19} = \infty$  for the normalization.

Finally, we add the following heaping pattern to the dependent variable: the duration points 4, 9, and 11 are rounded up to 5, 10, and 15 with probability 0.7, respectively, and the duration points 6, 11, and 16 are rounded down to 5, 10 and 15, respectively, with the same probability 0.7. Hence the heaping probability parameters are  $p_1 = q_1 = 0.7$ .

7. Please refer to ACG for details of the survey and the original sample used in ACG.

8.  $\exp(\kappa(t)) = \exp(-\delta(t)) = \exp(\gamma(0)) + \dots + \exp(\gamma(t-1))$ .

Algebraically, the actual observed duration variable `duration` is generated by:

$$\begin{aligned}
 u_i &\sim \text{Uniform}[0, 1] \\
 \text{duration}_i &= 5 \text{ if } \text{duration}_i = 4 \text{ and } u_i < 0.7 \\
 \text{duration}_i &= 5 \text{ if } \text{duration}_i = 6 \text{ and } u_i < 0.7 \\
 \text{duration}_i &= 10 \text{ if } \text{duration}_i = 9 \text{ and } u_i < 0.7 \\
 \text{duration}_i &= 10 \text{ if } \text{duration}_i = 11 \text{ and } u_i < 0.7 \\
 \text{duration}_i &= 15 \text{ if } \text{duration}_i = 14 \text{ and } u_i < 0.7 \\
 \text{duration}_i &= 15 \text{ if } \text{duration}_i = 16 \text{ and } u_i < 0.7
 \end{aligned}$$

We have not included the unobserved heterogeneity in the generation of the above data. Figure 2 plots the histograms of both observed duration variable with heaping and the true duration variable without heaping as generated from the ordered probit model.

## 5.2 `heapmph` command

This section describes the implementation of the `heapmph` command for the mixed proportional hazard model.

### Basic syntax

The basic syntax of the `heapmph` command follows the standard `Stata` command form:

```
heapmph depvar varlist [if] [in] [, options]
```

where `depvar` stands for the dependent variable, and `varlist` may contain the specified covariates. In this paper, we demonstrate the usages of the `heap` package with examples, and then explain a few other options available. We do not provide an exhaustive explanation of all the options available, and thus refer the interested user to the help files included in the package.

### Model estimation

As discussed in Section 5.1, the analysis is restricted to modeling the hazard rate during the first 18 days after birth since the reported number of deaths is smaller after this period (see ACG). We, therefore, add the `censor(18)` option to the command to fix the right-censoring period for each observation at 18. By default, the `heap` command assumes that the right-censoring period is the largest value of the dependent variable in the chosen sample. Instead of using the fixed right-censoring, it is also possible to allow for person-specific censoring points for each observation (see Section 5.4). We also provide a command to test for policy effects (see Appendix 2)

We next detail the values used for the four *compulsory* options to define the pattern of heaping in our example.

1. Since we have generated the data with heaps at days 5, 10, and 15, we define the starting period ( $h^*$ ) of 5 using the option `hstar(5)`. The assumption is that the heaping occurs at points that are multiples of  $h^*$ .
2. We set option `jbar(3)` (i.e.,  $\bar{j} = 3$ ) to indicate that there are a *maximum* of three heaping points prior to the censoring point (see point 1 above).
3. As illustrated in our stylized example (Figure 1), the rounding probabilities are  $p_1$ , and  $q_1$ , respectively. Hence, with the number of heaping probabilities, we have the maximum number of time periods that a duration can be rounded to is denoted as  $\bar{r} = 1$ . This is set by the option `rbar(1)` in the command.
4. The constant part of the baseline hazard enables us to identify the parameters of the heaping process. In this example, we set the time period after which the hazard is constant equal to 12 ( $\bar{k}$ ). Also, we assume that the heaping is asymmetric, which suggests that constant baseline hazard parameters are at different levels for periods  $\{12, 13, 14, 15\}$ .<sup>9</sup> In the command, the starting period of the flat segment can be defined by adding the option `kbar(12)`.

### ► Example

We choose `duration` as the dependent variable, and `age_m` and `school_m` as the covariates. We request Stata to implement the command using the code:

```
. heapmph duration age_m school_m, censor(18) hstar(5) jbar(3) kbar(12) rbar(1)
```

Coefficients estimation in progress (% finished approx.): 0%....1%.....10%  
.....20%.....30%.....40%.....50%  
.....60%.....70%.....80%.....90%  
.....100%

---

Initial temperature:	1	Final temperature:	0.000000010
Consecutive rejections:	10	Number of function calls:	35,277
Total final loss:	626.285	Observations:	250

---

MooN bootstrap will take approximately 25 minutes (100 replicates).  
(each dot . indicates one replication)

```

|-----| 1 |-----| 2 |-----| 3 |-----| 4 |-----| 5
..... 50
..... 100

```

---

	Coef.	Bootstrap Std. Err.	z	Std. Normal P> z	Bootstrap [95% Conf. Interval]	
exp(gamma)						
gamma0	.3057518	.0735412	4.16	0.000	.2955595	.3159441

9. See Assumption H (iii) in ACG.

	gamma1	.1539733	.0485058	3.17	0.002	.1472508	.1606959
	gamma2	.2757727	.0833442	3.31	0.001	.2642218	.2873236
	gamma3	.244938	.0663895	3.69	0.000	.2357369	.2541391
	gamma4	.2434391	.3819543	0.64	0.524	.1905029	.2963752
	gamma5	.5821127	.517579	1.12	0.261	.5103799	.6538455
	gamma6	.5240925	.9355118	0.56	0.575	.3944372	.6537478
	gamma7	.5396087	.1250595	4.31	0.000	.5222763	.556941
	gamma8	1.012341	.2346417	4.31	0.000	.9798218	1.044861
	gamma9	.6691035	1.008273	0.66	0.507	.529364	.808843
	gamma10	1.239431	.7855968	1.58	0.115	1.130553	1.348309
	gamma11	1.046086	1.855192	0.56	0.573	.7889691	1.303202
	gamma12	1.73946	.7507658	2.32	0.021	1.635409	1.843511
	gamma13	3.974719	5.473584	0.73	0.468	3.216119	4.733319
	gamma14	6.835998	2.162892	3.16	0.002	6.536237	7.13576
sigma	sigma	.0000838	.0041016	0.02	0.984	-.0004847	.0006522
beta	age_m	-.087153	.0098087	-8.89	0.000	-.0885124	-.0857936
	school_m	.1161123	.0108713	10.68	0.000	.1146056	.117619
prob_left	p1	.6780751	.3751911	1.81	0.071	.6260763	.730074
prob_right	q1	.6848905	.9971728	0.69	0.492	.5466894	.8230916

The command firstly employs a single simulated annealing algorithm (see Section 5.5.3) to solve for the point estimates. The  $M$  out of  $N$  bootstrap procedure is then conducted to yield the standard errors. Note also that the 95% bootstrap confidence interval is constructed using the 2.5% and the 97.5% quantile of the empirical bootstrap distribution. The output table consists of five panels. The panel **exp(gamma)** reports the estimates of functions of the baseline hazard parameters (see Section 5.1 and Appendix 3). It is worth mentioning again that we set the baseline hazard parameters  $\gamma_i$  to be constant over periods  $\{12, 13, 14, 15\}$ . Hence, the number of baseline hazard parameters we estimate is  $18 - 3 - 1 = 14$ . Specifically, **gamma0**, **gamma1**, ..., **gamma11** in the output table correspond to functions of the baseline hazard in period 0, 1, ..., 11, respectively. **gamma12** corresponds to the flat baseline hazard during periods  $\{12, 13, 14, 15\}$ . **gamma13** is for period 16, and **gamma14** is for the period 17.

Panel **sigma** displays the estimate of  $\sigma$  which is the standard deviation of the gamma distributed unobserved heterogeneity variable  $v_i$ , and panel **beta** is for the estimates of the covariate coefficients. In panels **prob\_left** and **prob\_right**, we report the estimated heaping probabilities  $p_1$ , and  $q_1$ . The value of **sigma** coefficient can be seen to be very close to zero numerically. This does not come unexpected since the data generating process does not feature any unobserved heterogeneity.<sup>10</sup>

◀

10. To test this formally, note that this is a test for a parameter on the boundary which requires an adjustment of the critical value or the p-value. Alternatively, for a formally valid likelihood ratio test, see Gutierrez et al. (2001).

### Testing for the presence of heaping effects

This command provides a subroutine to test null hypothesis via the Likelihood Ratio (LR) test described in Remark 4.2 in Section 4 of ACG, and briefly discussed in Section 4 in this paper. We provide a test (`testpi1`) that can be implemented by addition of an option to the main command. `testpi1` tests the null hypothesis ( $\mathbf{H}_0^{\pi_1}$ ) that all heaping probability parameters are zero, and the alternative ( $\mathbf{H}_A^{\pi_1}$ ) is that at least one heaping probability parameter is greater than zero. Applying the Intersection-Union principle (IUP), we could test the null hypothesis ( $\mathbf{H}_0^{\pi_2}$ ) that at least one heaping probability parameter is equal to zero, and the alternative ( $\mathbf{H}_A^{\pi_2}$ ) is that none is zero.

#### ► Example

To test for the presence of heaping effects under the model specification described in the last subsection, we can simply add `testpi1` option to the command:

```
. heapmph duration age_m school_m, censor(18) hstar(5) jbar(3) kbar(12) rbar(1)
testpi1
```

```
-- Estimating the coefficients under the null (H0) --
Coefficients estimation in progress (% finished approx.): 0%....1%.....10%
.....20%.....30%.....40%.....50%
.....60%.....70%.....80%.....90%
.....100%
-- Estimating the coefficients under the alternative (H1) --
Coefficients estimation in progress (% finished approx.): 0%....1%.....10%
.....20%.....30%.....40%.....50%
.....60%.....70%.....80%.....90%
.....100%
```

---

```
Moon bootstrap will take approximately 40 minutes (100 replications).
(each dot . indicates one replication)
|-----| 1 |-----| 2 |-----| 3 |-----| 4 |-----| 5
..... 50
..... 100
```

```
H0: all heaping probability parameters are zero
H1: at least one heaping probability parameters is greater than zero
```

QLR Statistic	[ The Bootstrap Critical Values ]		
	10%	5%	1%
24.981152	25.4834	25.5324	26.1653

The Stata output table reports the test statistic along with the corresponding bootstrapped critical values at 10%, 5% and 1% levels.<sup>11</sup> In this example, we fail to reject the null hypothesis at the 10% significance level, which suggests that there is no clear evidence of heaping.

◄

In addition, we employ the IUP rule to test the null that at least one heaping

11. `heapmph` stores 1<sup>st</sup>, 5<sup>th</sup>, 10<sup>th</sup>, 90<sup>th</sup>, 95<sup>th</sup> and 99<sup>th</sup> percentiles of the bootstrap empirical distribution function in `e()`. See the help file to this command for details.

probability parameter is equal to zero ( $\mathbf{H}_0^{\pi_2}$ ). In detail, we sort the p-values of all heaping parameters ( $p_1$  and  $q_1$ ) displayed in the regression output. The largest p-value is 0.492 in our example, so we do not reject the null at any conventional significance level, hence we have to continue to use  $M$  out of  $N$  subsampling scheme. Otherwise, if the null hypothesis was rejected, one could simply do inference based on the standard normal distribution.

### 5.3 heapop command for the ordered probit model with heaping

#### Basic syntax

The syntax and corresponding options of Stata command `heapop` are identical to those of the `heapmph` command (see Section 5.2).

#### Model estimation

The `heapop` command estimates an ordered probit model with heaping, and can be also employed to deal with the duration outcome data. The `heapop` command requires also four compulsory options to define the pattern of heaping, i.e., `kbar()`, `jbar()`, `hstar()`, and `rbar()`, as introduced in Section 5.2 for the `heapmph` command. In the case of ordered choice or count data, the `sensor()` option can be used to indicate the maximum number of possible choices or counts. If `sensor()` is left unspecified, Stata by default uses the maximum value of the dependent variable as `sensor()`.

This section attaches example usages of the `heapop` command under the same specification of the heaping pattern as used in Section 5.2.

We first request Stata to implement the `heapop` command to estimate the model:

#### ► Example

```
. heapop duration age_m school_m, hstar(5) jbar(3) kbar(12) rbar(1)
```

```
Coefficients estimation in progress (% finished approx.): 0%....1%.....10%
.....20%.....30%.....40%.....50%
.....60%.....70%.....80%.....90%
.....100%
```

Initial temperature:	1	Final temperature:	0.000000010
Consecutive rejections:	45	Number of function calls:	34,215
Total final loss:	593.206	Observations:	250

```
MooN bootstrap will take approximately 23 minutes (100 replicates).
(each dot . indicates one replication)
```

```

|-----| 1 |-----| 2 |-----| 3 |-----| 4 |-----| 5
.....50
.....100
```

---

Bootstrap

Std. Normal

Bootstrap



	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]	
exp(gamma)						
gamma0	.2721881	.075306	3.61	0.000	.2617512	.282625
gamma1	.3747939	.0963867	3.89	0.000	.3614354	.3881524
gamma2	.4633893	.1222535	3.79	0.000	.4464458	.4803327
gamma3	.4146261	.1156667	3.58	0.000	.3985955	.4306567
gamma4	.3467341	.6549929	0.53	0.597	.2559567	.4375114
gamma5	.5220416	.3889637	1.34	0.180	.4681339	.5759492
gamma6	.4701779	.9988004	0.47	0.638	.3317512	.6086046
gamma7	.3721548	.1083006	3.44	0.001	.3571451	.3871645
gamma8	.6864175	.1918713	3.58	0.000	.6598255	.7130095
gamma9	.694045	1.225619	0.57	0.571	.5241829	.8639071
gamma10	.6801856	.5940031	1.15	0.252	.597861	.7625103
gamma11	.3868165	.8571374	0.45	0.652	.2680233	.5056097
gamma12	.8342315	.5854894	1.42	0.154	.7530868	.9153762
gamma13	1.68294	2.923839	0.58	0.565	1.277717	2.088163
gamma14	1.991177	.6263514	3.18	0.001	1.90437	2.077985
beta						
age_m	-.0854586	.010742	-7.96	0.000	-.0869474	-.0839699
school_m	.0968189	.0027978	34.61	0.000	.0964311	.0972066
prob_left						
p1	.7011283	.4484472	1.56	0.118	.6389767	.7632799
prob_right						
q1	.6718045	1.121392	0.60	0.549	.5163875	.8272215

◀

Unlike the table in Section 5.2, this table consists of only four panels as no unobserved heterogeneity parameter has been estimated. Standard errors and bootstrap confidence intervals are constructed as before in Section 5.2. The first panel contains again the estimated baseline parameters (**exp(gamma)**; cf. Section 5.2 for the specification), while panel two provides estimates of the  $\beta$  coefficients. Note that the numerical differences in the  $\beta$  coefficient estimates is likely to stem from the omission of unobserved heterogeneity and the different functional form in this specification. Finally, panel three and four contain the estimated heaping probabilities, which can both be seen to be statistically insignificant.

### Testing for the presence of heaping effects

#### ► Example

To test for the presence of heaping effects ( $H^{\pi_1}$ ), we code:

```
. heapop duration age_m school_m, hstar(5) jbar(3) kbar(12) rbar(1) testpi1

-- Estimating the coefficients under the null (H0) --
Coefficients estimation in progress (% finished approx.): 0%....1%.....10%
.....20%.....30%.....40%.....50%
.....60%.....70%.....80%.....90%
.....100%
```

```
-- Estimating the coefficients under the alternative (H1) --
Coefficients estimation in progress (% finished approx.): 0%...1%.....10%
.....20%.....30%.....40%.....50%
.....60%.....70%.....80%.....90%
.....100%
```

```
MooN bootstrap will take approximately 41 minutes (100 replications).
(each dot . indicates one replication)
```

```
-----| 1 -----| 2 -----| 3 -----| 4 -----| 5
.....
..... 50
..... 100
```

H0: all heaping probability parameters are zero

H1: at least one heaping probability parameters is greater than zero

QLR Statistic	[ The Bootstrap Critical Values ]		
	10%	5%	1%
25.995373	26.2054	26.2146	26.2384

◀

As in the previous section, we cannot reject the null  $\mathbf{H}_0^{\pi_1}$  at any conventional level and thus proceed to test  $\mathbf{H}_0^{\pi_2}$  via the IUP rule. More specifically, we sort again the p-values of all heaping parameters ( $p_1$  and  $q_1$ ) displayed in the regression output. Since the largest p-value is 0.826, we do not reject the null at any conventional significance level and continue to use  $M$  out of  $N$  subsampling for inference.

## 5.4 Further options

Here, we elaborate on a few additional options, which are available for both commands `heapmph` and `heapop`.

### Bootstrap options

The `rep(integer)` option allows users to specify the number of  $M$  out of  $N$  bootstrap replications for calculating the standard errors. The default value is set at 100. In the example shown in Section 5.2, it takes 24 minutes to run 100 bootstrap iterations in a 64 bit Stata 15 SE on a desktop computer with the Intel i7 quad-core processor with 4.0GHz.

When choosing the  $M$  in the  $M$  out of  $N$  bootstrap, users can set the option `moon(real)` to select the share of  $M$  observations to be randomly drawn from the sample of size  $N$ . Bickel and Sakov (2008) provide an in-depth discussion on the choice of the  $M$  parameter. The `heap` packages, by default, set `moon` at 0.8 so that in each MooN bootstrap iteration, 80% of the original sample are randomly kept.

## Optimization

The commands provided implements the Simulated Annealing (SA) algorithm to maximize the likelihood function of the model. The SA method, proposed by Kirkpatrick et al. (1983), is a popular local search algorithm for stochastically approximating the global optimum of a given objective function. The review of the algorithm and its technical details can be found in Dowsland and Thompson (2012), for example. The SA algorithm is particularly useful for our model, and may be preferable to the conventional Newton algorithm, since SA is better at locating global maximum when the likelihood function is complex, as in our case.

The `heap` package self-contains the `Mata` function for SA method of Kirkpatrick et al. (1983). In this function, we have designed 10 options for users to control settings of the SA algorithm. For instance, `sa_verbosity(integer)` allows the user to set the maximum number of total iterations (the default is 8000) and the `sa_stopTemp(real)` option allows one to set the temperature at which to stop the searching algorithm (the default is  $1 \times 10^{-8}$ ). The full details about the settings are listed in the help file to this command. Besides, the seed state for initializing the random number generator is set to be 1000 by default, and can be adjusted in the `seed(real)` option.<sup>12</sup>

## Display options

For diagnosing and monitoring purposes, we provide the following two options to display the intermediate command outputs. First, the `detail` option can be used to display a summary of heaping model specifications, and produce a table only for point estimates before conducting the bootstrap. Second, the `sa_verbosity(integer)` (**DG: is this correct?**) option can be set to 1 for producing the final report of the simulated annealing, and set to 2 for further displaying the temperature changes in each iteration. The default value of this option is zero which suppresses all output.

## Different censoring points for each observation

The option for variable censoring is `vcensor(varname)`, where *varname* is a dummy variable which equals to 1 if the observation is complete and is 0 if the observation is right-censored.

Let `uncensor_dummy` stand for a period-specific censoring indicator variable. `uncensor_dummy=1` if the observation's spell is complete, and `uncensor_dummy=0` if the spell is right-censored. For example, we randomly generate `uncensor_dummy` from a Bernoulli(0.1) distribution, and apply the `heapmph` command:

```
. generate byte uncensor_dummy = uniform() <0.1
```

12. Another user-written `Mata` function is 'simann'. We have not used this since, we did not know how the function actually performed as the author did not disclose the source code of this function. Additionally, the command was not flexible enough, since some of the parameters were fixed in the 'simann' function. Based on the Matlab's simulated annealing function, one of the authors (Zizhong Yan) has programmed a more flexible `Mata` simulated annealing function for our heaping command.

```
. heapmph duration age_m school_m,vcensor(uncensor_dummy) hstar(5) jbar(3) kbar(12)  
rbar(2)
```

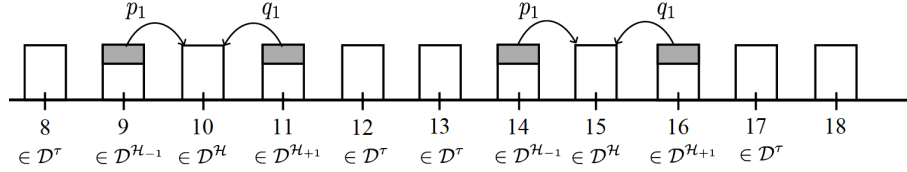
(output omitted)

Note that if neither `vcensor(varname)` nor `ccensor(integer)` is specified, the command by default will fix the right-censoring point at the maximum value of the dependent variable in the usable sample.

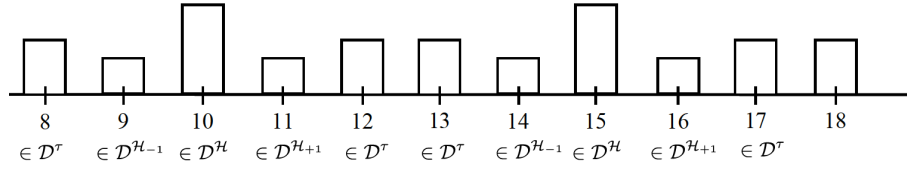
## 6 Conclusion

Discrete time duration models are very popular among researchers. The **Stata** command `heapmph` allows the estimation of a discrete time mixed proportional hazard model, when the observed discrete durations exhibit abnormal concentrations at certain durations points. An accompanying code `heapop` allows for heaping in an ordered probit model. The underlying assumptions and the identification strategy used are discussed fully in ACG.

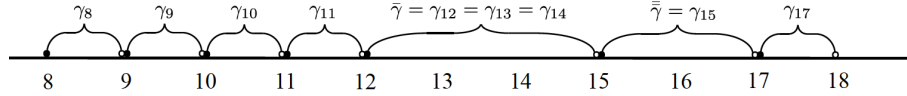
Figure 1: Stylized Example  
A: Heaping pattern



B: Observed data

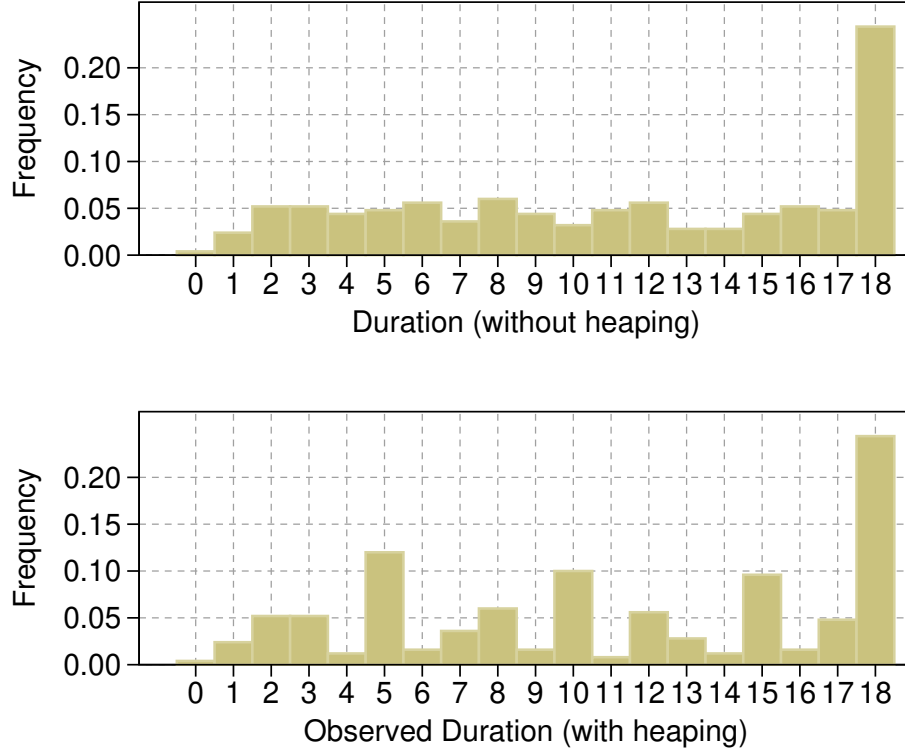


C: Baseline hazard



Notes: (i) This stylized example allows the heaps at periods 10 and 15. (ii)  $\mathcal{D}^H$  is the set of the reported durations on the heaping points.  $\mathcal{D}^{H-1}$  stands for the set reported duration points that are one period below the nearest heaping point. Similarly,  $\mathcal{D}^{H+1}$  stands for the set of reported durations that are one period above the nearest heaping point.  $\mathcal{D}^r$  refers to the correctly reported durations. (iii) The rounding probabilities of heaping are  $p_1$  and  $q_1$  for  $\mathcal{D}^{H-1}$  and  $\mathcal{D}^{H+1}$ , respectively. (iv) The constant part of the baseline hazard starts from period 12. By the asymmetric heaping, the constant baseline hazard parameters ( $\gamma_t = \ln \int_t^{t+1} \lambda_0(s) ds$ ) are at different levels for periods [12, 15) and [15, 17). (v) In Stata output table in Section 5.2, **gamma8**, **gamma9**, ..., **gamma11** correspond to the baseline hazard in period 8, 9, ..., 11, respectively. **gamma12** corresponds to the constant baseline hazard during periods [12, 15). **gamma13** is for period [15, 17) and **gamma14** is for the period 17. The period 18 in this example is the right-censoring date.

Figure 2: Histograms of the duration variable in the example data for demonstrating `heapmph` command (See Section 5.1)



Notes: (i) The upper graph plots the unobserved true duration variable without the heaping pattern. (ii) The lower graph presents the observed duration variable. In this example, duration points 4, 9, and 11 are rounded up to 5, 10, and 15 with probability 0.7, respectively. Duration points 6, 11, and 16 are rounded down to 5, 10, and 15 with the same probability 0.7, respectively. (i.e.,  $p_1 = q_1 = 0.7$ ) (iii) The right-censoring date is the period 18 in this data.

Table 1: Summary statistics for the variables used in the illustration of `heapmph` command

Variable	Mean (SD)
Number of days of survival of the children excluding the censored observations	8.873 (4.828)
Proportion of censored observations at 18 days	0.244 (0.430)
Age of mother at the birth of the child, in years	24.060 (5.120)
Mother's education, in years	3.248 (4.135)
Proportion of children who were born during the treatment period	0.132 (0.339)
Total number of children	250

Notes: See Section 5.1 for the model that generated this data.

## 7 References

- Abbring, J. H., and G. J. Van Den Berg. 2007. The unobserved heterogeneity distribution in duration analysis. *Biometrika* 94(1): 87–99.
- Andrews, D. W. 2000. Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica* 68(2): 399–405.
- Arulampalam, W., V. Corradi, and D. Gutknecht. 2017. Modeling heaped duration data: An application to neonatal mortality. *Journal of Econometrics* 200(2): 363–377.
- Bickel, P. J., and A. Sakov. 2008. On the choice of  $m$  in the  $m$  out of  $n$  bootstrap and confidence bounds for extrema. *Statistica Sinica* 18(3): 967–985.
- Cox, D. R. 1972. Models and life-tables regression. *Journal of Royal Statistical Society: Series B* 34: 187–220.
- Cox, D. R., and D. Oakes. 1984. Analysis of survival data. *Chapman&Hall, London*.
- Dowsland, K. A., and J. M. Thompson. 2012. Simulated annealing. In *Handbook of Natural Computing*, 1623–1655. Springer.
- Forster, M., and A. M. Jones. 2001. The role of tobacco taxes in starting and quitting smoking: duration analysis of British data. *Journal of the Royal Statistical Society: Series A* 164(3): 517–547.
- Greene, W. 2014. Models for Ordered Choices. In *Handbook of Choice Modelling*, ed. S. Hess and A. Daly, 333–362. Edward Elgar Publishing.

- Gutierrez, R. G., S. Carter, and D. M. Drukker. 2001. On boundary-value likelihood-ratio tests. *Stata Technical Bulletin* 10(60).
- Han, A., and J. A. Hausman. 1990. Flexible parametric estimation of duration and competing risk models. *Journal of applied Econometrics* 5(1): 1–28.
- Kirkpatrick, S., C. D. Gelatt, and M. P. Vecchi. 1983. Optimization by simulated annealing. *Science* 220(4598): 671–680.
- Meyer, B. D. 1990. Unemployment insurance and unemployment spells. *Econometrica* 58(4): 757–782.
- Pudney, S. 2008. Heaping and leaping: Survey response behaviour and the dynamics of self-reported consumption expenditure. Technical report, ISER Working Paper Series.
- Silvapulle, M., and P. Sen. 2005. *Constrained Statistical Inference: Inequality, Order, and Shape Restrictions* (Wiley Series in Probability and Statistics).

**About the authors**

Wiji Arulampalam, Department of Economics, University of Warwick, Coventry CV4 7AL, UK. Email: Wiji.Arulampalam@warwick.ac.uk

Valentina Corradi, Department of Economics, University of Surrey, School of Economics, Guildford GU2 7XH, UK. Email: V.Corradi@surrey.ac.uk

Daniel Gutknecht, Department of Economics, University of Mannheim, L7, 3-5, 68131 Mannheim, Germany. Email: Daniel.Gutknecht@gmx.de

Zizhong Yan (Corresponding Author), Institute for Economic and Social Research, Jinan University, Guangzhou, China. Email: helloyzz@gmail.com



## **Appendix to the paper**

## 1 Estimating the policy effect

In this Appendix, we discuss an additional feature of the command `heapmph` and `heapop`, that allows the user to test for a shift in the baseline hazard in the duration model or the threshold parameters in the ordered probit model, and/or the reporting probabilities, perhaps due to a change in a binary variable.<sup>13</sup> For example, one might be interested in the analysis of the effects of a certain policy change on duration outcomes, and the binary indicator will then take the value of one for treated individuals. ACG's main focus, for example, is on whether the Janani Siraksha Yojana (JSY) program in India had any effect on neo-natal mortality, as well as on women's reporting behavior. The hypothesis being that more accurate records are available on average, compared to before, as the program encouraged women to deliver babies in health facilities.

The `treat(varname)` option of `heapmph` and `heapop` commands allows the user to account for the effect of a policy change on duration outcomes where `varname` is the name of the binary indicator variable. The treatment indicator variable is the actual treatment status for the 250 children randomly chosen from the original ACG data set, and as reported in Table 1, 13.2% of the children in our sample, were born during the treatment period. Since the data set used here are the same as that discussed in Section 5.2, we would expect to not reject the null hypothesis of zero treatment effects on the gamma parameters and the misreporting probabilities.

### ► Example

In the data example used in this paper, the `jsy_dummy` variable is the indicator for whether the JSY program was in place at the time of birth of the child. Taking the example of `heapmph` command, we code:

```
. heapmph duration age_m school_m, treat(jsy_dummy) censor(18) hstar(5) jbar(3)
kbar(12) rbar(1)
```

Coefficients estimation in progress (% finished approx.): 0%....1%.....10%  
.....20%.....30%.....40%.....50%  
.....60%.....70%.....80%.....90%  
.....100%

---

Initial temperature:	1	Final temperature:	0.000000010
Consecutive rejections:	252	Number of function calls:	14,911
Total final loss:	509.181	Observations:	250

---

MooN bootstrap will take approximately 15 minutes (100 replicates).  
(each dot . indicates one replication)

```

|-----| 1 |-----| 2 |-----| 3 |-----| 4 |-----| 5
.....
..... 50
..... 100
```

---

	Coef.	Bootstrap Std. Err.	z	Std. Normal P> z	Bootstrap [95% Conf. Interval]
--	-------	------------------------	---	---------------------	-----------------------------------

13. Similar tests can also be carried out in the ordered probit model, where the treatment is allowed to shift the gamma parameters and also the mis-reporting probabilities. In order to save space, we do not report an example using the `heapop` command here.

exp(gamma)						
gamma0	.1229606	.4263486	0.29	0.773	.0638716	.1820495
gamma1	.0386019	.2927666	0.13	0.895	-.0019735	.0791773
gamma2	.1268302	.3835026	0.33	0.741	.0736794	.1799809
gamma3	.1383078	.3970532	0.35	0.728	.083279	.1933366
gamma4	.0875479	.5449995	0.16	0.872	.0120148	.1630809
gamma5	.0173866	2.212301	0.01	0.994	-.2892226	.3239959
gamma6	.5470309	2.963335	0.18	0.854	.1363337	.9577282
gamma7	.3435419	.733828	0.47	0.640	.2418385	.4452453
gamma8	.6535276	1.153605	0.57	0.571	.4936461	.813409
gamma9	.2636071	2.474182	0.11	0.915	-.0792969	.6065112
gamma10	.2316901	3.614136	0.06	0.949	-.2692036	.7325839
gamma11	1.36546	4.803913	0.28	0.776	.6996713	2.031248
gamma12	1.093716	2.02697	0.54	0.589	.8127926	1.37464
gamma13	3.628725	10.84573	0.33	0.738	2.125584	5.131866
gamma14	12.31933	16.26165	0.76	0.449	10.06558	14.57308
exp(gamma_tr)						
gamma_treat0	1.756707	1.980838	0.89	0.375	1.482177	2.031237
gamma_treat1	1.315527	2.658768	0.49	0.621	.9470402	1.684013
gamma_treat2	1.342548	1.891511	0.71	0.478	1.080398	1.604698
gamma_treat3	.9620915	2.02116	0.48	0.634	.681973	1.24221
gamma_treat4	1.401458	3.599703	0.39	0.697	.9025645	1.900351
gamma_treat5	1.691805	3.873919	0.44	0.662	1.154907	2.228703
gamma_treat6	1.808029	3.684712	0.49	0.624	1.297354	2.318704
gamma_treat7	.5005411	1.533699	0.33	0.744	.2879813	.7131009
gamma_treat8	.6602585	1.314985	0.50	0.616	.4780109	.8425061
gamma_treat9	.0465493	6.273625	0.01	0.994	-.8229308	.9160294
gamma_tre-10	.5768342	4.701807	0.12	0.902	-.0748031	1.228471
gamma_tre-11	.8627947	3.274777	0.26	0.792	.4089338	1.316656
gamma_tre-12	1.450397	2.087993	0.69	0.487	1.161016	1.739778
gamma_tre-13	-.4097875	5.926568	-0.07	0.945	-1.231168	.411593
gamma_tre-14	-2.876843	3.571297	-0.81	0.421	-3.371799	-2.381886
sigma						
sigma	.1412589	.834647	0.17	0.866	.0255827	.2569351
beta						
age_m	-.0607541	.1016391	-0.60	0.550	-.0748406	-.0466677
school_m	.1220928	.24783	0.49	0.622	.0877453	.1564403
prob_left						
p1	.5884231	1.075745	0.55	0.584	.4393325	.7375137
prob_right						
q1	.8144293	2.117911	0.38	0.701	.5209018	1.107957
prob_left_treat						
p1D	-.3078594	2.353369	-0.13	0.896	-.6340198	.0183009
prob_right_tr-t						
q1D	.9277816	.4917421	1.89	0.059	.8596296	.9959336

The specifications of the heaping pattern is same as the one in Section 5. This Stata output table has the same format as the output table in Section 5.2. In particular, the panel `exp(gamma_treat)` in this table reports the estimated baseline parameters for the treatment group units (i.e.,  $\exp(\gamma^{(2)}(1))$ ). Panels `prob_left_treat`

presents the estimated change of the heaping probabilities ( $p_1^{(2)}$ ) of the treatment group. `prob_left_right` reports ( $q_1^{(2)}$ ) of the treatment group.

◀

### Testing hypotheses

When estimating the policy effect, the `heapmph` command provides two options for testing hypotheses as follows.

#### Test for the changes in the reporting behavior after the policy introduction

As outlined in Section 5 of ACG, first we would like to rule out that changes in the reporting behavior (as a result of the policy introduction) confound any observable effect of the program. Therefore, we start by testing  $\mathbf{H}^{\pi_3}$ , which under the null ( $\mathbf{H}_0^{\pi_3}$ ) postulates that all deviations  $p_1^{(2)}$  and  $q_1^{(2)}$  are jointly equal to zero.

#### ► Example

For instance, we could use the `testpi3` option:

```
. heapmph duration age_m school_m, treat(jsy_dummy) censor(18) hstar(5) jbar(3)
kbar(12) rbar(1) testpi3
```

```
-- Estimating the coefficients under the null (H0) --
Coefficients estimation in progress (% finished approx.): 0%...1%.....10%
.....20%.....30%.....40%.....50%
.....60%.....70%.....80%.....90%
.....100%
-- Estimating the coefficients under the alternative (H1) --
Coefficients estimation in progress (% finished approx.): 0%...1%.....10%
.....20%.....30%.....40%.....50%
.....60%.....70%.....80%.....90%
.....100%
```

---

```
MooN bootstrap will take approximately 36 minutes (100 replications).
(each dot . indicates one replication)
----- 1 ----- 2 ----- 3 ----- 4 ----- 5
..... 50
..... 100
```

```
H0: treatment has not changed the exit probability
H1: over at least one period the exit probability decreased
```

QLR Statistic	[ The Bootstrap Critical Values ]		
	10%	5%	1%
22.797652	115.5279	160.6586	276.2059

Here, in this illustration, we cannot reject the null hypothesis that there is no change in the heaping probability parameters after the policy introduction, at the 10% level.

◀

### Test for whether the treatment has changed the exit probability

The `heapmph` command provides the `testgamma2` option to test for the null hypothesis ( $\mathbf{H}_0^{72}$ ) that treatment has not changed the exit probability (e.g., the probability of the event happens) in any of the first  $(\bar{\tau} - 1)$  periods against the alternative ( $\mathbf{H}_A^{72}$ ) that over at least one period the exit probability decreased. For the technical details of this test, see Section 5 of ACG.

#### ► Example

In Stata, we code:

```
. heapmph duration age_m school_m, treat(jsy_dummy) censor(18) hstar(5) jbar(3)
kbar(12) rbar(1) testgamma2
```

```
-- Estimating the coefficients under the null (H0) --
Coefficients estimation in progress (% finished approx.): 0%...1%.....10%
.....20%.....30%.....40%.....50%
.....60%.....70%.....80%.....90%
.....100%
-- Estimating the coefficients under the alternative (H1) --
Coefficients estimation in progress (% finished approx.): 0%...1%.....10%
.....20%.....30%.....40%.....50%
.....60%.....70%.....80%.....90%
.....100%
```

---

```
MooN bootstrap will take approximately 26 minutes (100 replications).
(each dot . indicates one replication)
|-----| 1 |-----| 2 |-----| 3 |-----| 4 |-----| 5
..... 50
..... 100
```

```
H0: no change in the heaping probability parameters after the policy (treatment) introduction
H1: a change in at least some rounding parameters
```

QLR Statistic	[ The Bootstrap Critical Values ]		
	10%	5%	1%
101.96437	241.9557	272.1655	306.3246

From the output tables, we find that the null of  $\mathbf{H}_0^{72}$  cannot be rejected at a 10% significance level.

◀

## 2 Policy analysis: using heapop command

#### ► Example

One might be interested in using the ordered probit model to estimate the effects of a certain policy change on the heaping probabilities. We code:

```
. heapop duration age_m school_m, treat(jsy_dummy) hstar(5) jbar(3) kbar(12) rbar(1)
```

Coefficients estimation in progress (% finished approx.): 0%....1%.....10%  
 .....20%.....30%.....40%.....50%  
 .....60%.....70%.....80%.....90%  
 .....100%

Initial temperature:	1	Final temperature:	0.000000010
Consecutive rejections:	0	Number of function calls:	17,079
Total final loss:	580.834	Observations:	250

MooN bootstrap will take approximately 17 minutes (100 replicates).  
 (each dot . indicates one replication)

.....	1	.....	2	.....	3	.....	4	.....	5
.....		.....		.....		.....		.....	
.....		.....		.....		.....		.....	

	Coef.	Bootstrap Std. Err.	z	Std. Normal P> z	Bootstrap [95% Conf. Interval]	
exp(gamma)						
gamma0	.2201253	.1222564	1.80	0.072	.2031814	.2370692
gamma1	.3234947	.1733278	1.87	0.062	.2994727	.3475167
gamma2	.373393	.2060417	1.81	0.070	.3448371	.4019489
gamma3	.2748416	.1529418	1.80	0.072	.2536449	.2960382
gamma4	.1865983	.5957995	0.31	0.754	.1040247	.2691719
gamma5	.4966012	.6241665	0.80	0.426	.4100962	.5831063
gamma6	.3905122	.940318	0.42	0.678	.2601907	.5208336
gamma7	.4948187	.3479875	1.42	0.155	.4465901	.5430473
gamma8	.5846685	.3741055	1.56	0.118	.5328202	.6365169
gamma9	.3797094	1.122963	0.34	0.735	.2240747	.5353441
gamma10	.6438094	.7899548	0.81	0.415	.5343273	.7532916
gamma11	.5438619	1.104397	0.49	0.622	.3908003	.6969236
gamma12	.619283	.9040273	0.69	0.493	.4939912	.7445748
gamma13	1.587706	4.084947	0.39	0.698	1.021561	2.153851
gamma14	1.736305	1.103575	1.57	0.116	1.583357	1.889252
exp(gamma_tr)						
gamma_treat0	-.0277171	3.423028	-0.01	0.994	-.5021246	.4466905
gamma_treat1	.4552177	2.410112	0.19	0.850	.1211932	.7892421
gamma_treat2	1.580163	1.246502	1.27	0.205	1.407406	1.752919
gamma_treat3	2.63764	1.641752	1.61	0.108	2.410105	2.865175
gamma_treat4	2.118725	2.607281	0.81	0.416	1.757375	2.480076
gamma_treat5	.1418619	1.802993	0.08	0.937	-.1080202	.391744
gamma_treat6	1.270889	2.182315	0.58	0.560	.9684354	1.573342
gamma_treat7	-2.791271	5.147388	-0.54	0.588	-3.504663	-2.07788
gamma_treat8	.4949781	.6722757	0.74	0.462	.4018055	.5881508
gamma_treat9	.2580867	5.021704	0.05	0.959	-.4378859	.9540594
gamma_tre-10	.8517118	3.613198	0.24	0.814	.350948	1.352476
gamma_tre-11	-.5838686	5.963755	-0.10	0.922	-1.410403	.2426656
gamma_tre-12	1.044385	.7260383	1.44	0.150	.9437617	1.145009
gamma_tre-13	.388965	3.01887	0.13	0.897	-.029429	.8073589
gamma_tre-14	.391813	.4705357	0.83	0.405	.3266001	.457026
beta						
age_m	-.075491	.0198171	-3.81	0.000	-.0782375	-.0727445
school_m	.0978651	.0102829	9.52	0.000	.09644	.0992903
prob_left						
p1	.5402499	.4846781	1.11	0.265	.4730769	.6074228

prob_right						
q1	.7557395	2.360927	0.32	0.749	.4285317	1.082947
prob_left_treat						
p1D	.1991267	1.050912	0.19	0.850	.0534778	.3447756
prob_right_tr-t						
q1D	-.5226667	2.121365	-0.25	0.805	-.816673	-.2286605

◀

Similar to the Appendix 1, the `heapop` command provides two options for testing hypotheses for the policy analysis.

### ► Example

First, to test for the changes in the reporting behavior after the policy introduction ( $H^{\pi_3}$ ), one can code:

```
. heapop duration age_m school_m, treat(jsy_dummy) hstar(5) jbar(3) kbar(12) rbar(1)
testpi3

-- Estimating the coefficients under the null (H0) --
Coefficients estimation in progress (% finished approx.): 0%...1%.....10%
.....20%.....30%.....40%.....50%
.....60%.....70%.....80%.....90%
.....100%
-- Estimating the coefficients under the alternative (H1) --
Coefficients estimation in progress (% finished approx.): 0%...1%.....10%
.....20%.....30%.....40%.....50%
.....60%.....70%.....80%.....90%
.....100%
```

---

MooN bootstrap will take approximately 30 minutes (100 replications).  
(each dot . indicates one replication)

	1		2		3		4		5	
.....										50
.....										100

H0: treatment has not changed the exit probability  
H1: over at least one period the exit probability decreased

QLR Statistic	[ The Bootstrap Critical Values ]		
	10%	5%	1%
-2.3731951	1.9105	2.3122	4.7291

◀

### ► Example

Second, we test for whether the treatment has changed the exit probability ( $H^{\gamma_2}$ ):

```
. heapop duration age_m school_m, treat(jsy_dummy) hstar(5) jbar(3) kbar(12) rbar(1)
testgamma2
```





Here,  $\delta_\tau$  denotes the log of the integrated baseline hazard given by:

$$\delta_\tau = \ln \int_0^\tau \lambda_0(s) ds, \quad \tau = 1, \dots, \bar{\tau} - 1. \quad (7)$$

The relationship between these and the baseline hazard function parameters we saw earlier, is given by:

$$\begin{aligned} \exp(\delta_\tau) &= \int_0^\tau \lambda_0(\tau^*) d\tau^* = \int_0^1 \lambda_0(\tau^*) d\tau^* + \dots + \int_{\tau-1}^\tau \lambda_0(\tau^*) d\tau^* \\ &= \exp(\gamma(0)) + \exp(\gamma(1)) + \dots + \exp(\gamma(\tau-1)) \end{aligned}$$

We next replace the assumption for the distribution of  $\varepsilon_i$  in (4) with Type I EV distribution to obtain:

$$\begin{aligned} \Pr(\tau_i = \tau) &= \Pr(\kappa_\tau \leq y_i^* < \kappa_{\tau+1}) \\ &= \exp(-\exp(\kappa_{\tau+1} - z_i' \beta^\dagger)) - \exp(-\exp(\kappa_\tau - z_i' \beta^\dagger)) \end{aligned} \quad (8)$$

Comparing (6) with (8), we note that  $\beta = -\beta^\dagger$ . That is, a variable that has an effect of increasing the exit rate (hazard) will obviously have a decreasing effect on the duration. The ordered outcome model links neatly to the underlying hazard model in the case of the proportional hazard specifications. There is no one-to-one relationship between these two in other models such as the ordered probit or the ordered logit.

In summary, the ordered choice model derived by assuming a Type I EV distribution for the underlying latent variable equation error  $\varepsilon_i$ , is equivalent to the discrete duration model derived from a continuous time proportional hazard model.